Claims

1. A data quality system for matching input data across data records, the system comprising:-

means for pre-processing the input data to remove noise or reformat the data,

means for matching record pairs based on measuring similarity of selected field pairs within the record, and for generating a similarity indicator for each record pair.

2. A system as claimed in claim 1, wherein the matching means comprises means for extracting a similarity vector for each record pair by generating a similarity score for each of a plurality of pairs of fields in the records, the set of scores for a record pair being a vector.

3. A system as claimed in claim 2, wherein the vector extraction means comprises means for executing string matching routines on pre-selected field pairs of the records.

4. A system as claimed in claim 3, wherein a matching routine comprises means for determining an edit distance indicating the number of edits required to change from one value to the other value.

5. A system as claimed in claims 3 or 4, wherein a matching routine comprises means for comparing numerical values by applying numerical weights to digit positions.

6. A system as claimed in any of claims 2 to 5, wherein the vector extraction means comprises means for generating a vector value between 0 and 1 for each field pair in a record pair.

7. A system as claimed in any of claims 2 to 6, wherein the matching means comprises record scoring means for converting the vector into a single similarity score representing overall similarity of the fields in each record pair.

8. A system as claimed in claim 7, wherein the record scoring means comprises means for executing rule-based routines using weights applied to fields according to the extent to which each field is indicative of record matching.

9. A system as claimed in claims 7 or 8, wherein the record scoring means comprises means for computing scores using an artificial intelligence technique to deduce from examples given by the user an optimum routine for computing the score from the vector.

10. A system as claimed in claim 9, wherein the artificial intelligence technique used is cased based reasoning (CBR).

11. A system as claimed in claim 9, where the artificial intelligence technique used comprises neural network processing.

12. A system as claimed in any preceding claim, wherein the pre-processing means comprises a standardisation module comprising means for transforming each data field into one or more target data fields each of which is a variation of the original.

13. A system as claimed in claim 12, wherein the standardisation module comprises means for splitting a data field into multiple field elements, converting the field elements to a different format, removing noise characters, and replacing elements with equivalent elements selected from an equivalence table.

14. A system as claimed in any preceding claim, wherein the pre-processing means comprises a grouping module comprising means for grouping records according to features to ensure that all actual matches of a record are within a group, and wherein the matching means comprises means for comparing records within groups only.

15. A system as claimed in claim 14, wherein the grouping module comprises means for applying labels to a record in which a label is determined for a plurality of fields in a record and records are grouped according to similarity of the labels.

16. A system as claimed in claim 15, in which a label is a key letter for a field.

17. A system as claimed in any preceding claim, wherein the system further comprises a configuration manager comprising means for applying configurable settings for the pre-processing means and for the matching means.

18. A system as claimed in any of claims 7 to 17, wherein the system further comprises a tuning manager comprising means for refining, according to user inputs, operation of the record scoring means.

19. A system as claimed in claim 18, wherein the tuning manager comprises means for using a rule-based approach for a first training run and an artificial intelligence approach for subsequent training runs.